System Description for BioNNE-R at BioASQ 2026: OpenNRE with Pre-trained Language Models and Threshold Calibration for Biomedical Nested Relation Extraction

Notebook for the BioNNE-R Task of the BioASQ Lab at CLEF 2026

Savva Blinov[1,*]

[1]HSE University, Moscow, Russia

Abstract

We describe our system for the BioNNE-R shared task at BioASQ 2026, which addresses the extraction of semantic relations between nested named entity mentions in biomedical scientific abstracts in English and Russian. Our approach builds upon the OpenNRE framework with entity marker-based sentence encoding, using BioLinkBERT-large for the English track and XLM-RoBERTa-large for the Russian and bilingual tracks. We introduce three post-processing steps: entity-type masks, distance-based candidate pruning, and per-class decision threshold calibration in order to address the high false-positive rate arising from exhaustive candidate pair enumeration at inference time. For the English track, we additionally combine two models trained with different random seeds via logit averaging. All final models are trained on the combined train and development data. Our system achieves macro F1 scores of 0.4570 for English, 0.4829 for Russian, and 0.4849 for the bilingual track on the test set, outperforming the official OpenNRE baseline, whose corresponding scores are 0.2825, 0.3070, and 0.3027, by more than 0.17 F1 across all three tracks.

Keywords

relation extraction, biomedical NLP, nested named entities, multilingual NLP, BioASQ, NEREL-BIO

1. Introduction

Relation extraction (RE) is a fundamental task in biomedical natural language processing, enabling the automatic construction of knowledge graphs from scientific literature. The field has matured significantly over the past decade: early pattern-based and kernel methods have been largely supplanted by transformer-based fine-tuning approaches, which achieve strong results on established English benchmarks such as ChemProt [3] and BioRED [4]. Entity-marker methods, which insert special tokens around entity spans in the input sequence, have proven particularly effective for sentence-level RE [2], and domain-specific pre-trained models such as BioBERT [5] and PubMedBERT [6] provide further gains over general-purpose encoders for English biomedical text.

However, two important dimensions remain underexplored in the existing literature. First, the vast majority of biomedical RE systems assume flat entity mentions, where each entity occupies a single non-overlapping span. In practice, biomedical entities are often compositional and nested: in the phrase "atopic dermatitis", the inner entity "dermatitis" is nested inside the outer entity, and a SUBCLASS_OF relation holds between them. Standard pipelines are not designed to enumerate or classify such nested pairs. Second, most existing systems target English text exclusively. Russian biomedical literature is substantial but underserved: cross-lingual models such as XLM-RoBERTa [7] cover Russian, but there are few dedicated biomedical RE resources or systems for the language.

BioASQ is a long-running series of shared tasks promoting advances in large-scale biomedical semantic indexing and question answering [11]. Starting from BioASQ 2024, the challenge has been extended with tasks targeting biomedical named entity recognition and linking over nested mentions.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

blinov.sawa@gmail.com (S. Blinov)

The BioNNE task introduced in 2024 [12] focused on nested named entity recognition (NER) in Russian and English biomedical abstracts, with the top system, BINDER with XLM-RoBERTa, achieving macro F1 of 0.70 on the bilingual track. BioNNE-L 2025 [9] extended the challenge to entity linking normalising nested entity mentions to UMLS concept identifiers adding the complexity of multilingual terminology. BioNNE-R 2026 [1] represents a further step, shifting the focus from entity-level tasks to relation extraction between nested mentions, which requires reasoning about pairs of entities rather than individual spans.

The BioNNE-R shared task at BioASQ 2026 [1], part of the fourteenth BioASQ challenge [15], addresses both gaps simultaneously. It requires extracting relations between nested named entity mentions from biomedical PubMed abstracts in English and Russian, across three tracks: monolingual English, monolingual Russian, and bilingual. The evaluation metric is macro-averaged F1 over 14 relation types, putting equal weight on rare and frequent classes, a deliberate design choice that rewards systems capable of handling the full distribution of biomedical relations, not just the dominant ones.

In this paper, we describe our system, which extends the OpenNRE framework [2] with several domain-specific components. We demonstrate that entity-type masking, per-class threshold calibration, and distance-based pruning provide consistent improvements over a plain fine-tuned classifier, and that training on the full train+dev data further boosts test performance. BioLinkBERT [8], which additionally leverages inter-document citation link structure during pre-training on PubMed, serves as our English backbone. The predecessor task BioNNE-L 2025 [9] focused on entity linking rather than relation extraction, but shared the nested entity structure and multilingual NEREL-BIO data that our system builds upon.

2. Related Work

Neural relation extraction spans a spectrum of architectures. Sentence-level classifiers that insert entity-marker tokens around the head and tail spans and read out a fixed representation, typically the [CLS] token, are simple, fast, and strong on benchmarks such as ChemProt [3]; this is the design adopted by the OpenNRE toolkit [2]. At the other end of the spectrum, span-pair and graph-based models, which enumerate span representations and reason over document-level or dependency graphs, can capture long-range and cross-sentence interactions and are better suited to document-level datasets such as BioRED [4], at the cost of more parameters and more supervision. Because the BioNNE-R organisers release an OpenNRE entity-marker baseline [1] and the English track provides only 55 training documents, we deliberately build on the sentence-level marker design rather than a graph model: it is data-efficient and reproduces a well-understood baseline, which is an appropriate starting point in a low-resource, shared-task setting.

A recurring idea in structured prediction is to constrain the output space using the types of the arguments. In knowledge-base completion, ontological type constraints are used to prune impossible (subject, predicate, object) triples, and analogous schema constraints appear in information-extraction pipelines. Our entity-type masks are an instance of this idea, but the allowed (head_type, tail_type) sets are estimated empirically from the training annotations rather than imported from an external ontology, which keeps them consistent with the NEREL-BIO [10] annotation decisions and requires no additional resources.

Finally, our per-class threshold calibration is related to the confidence-calibration literature. Temperature scaling [14] and Platt scaling adjust a classifier's probabilities to better match empirical accuracy, typically by optimising a likelihood-based objective. In contrast, we tune one decision threshold per relation directly against the task's macro-F1 objective, which places equal weight on rare and frequent relations. Our low-resource observations for the English track are consistent with the broader finding that fine-tuning large pre-trained encoders is unstable and high-variance across random seeds [13].

3. Task and Data Description

3.1. Task Overview

The BioNNE-R task uses data derived from NEREL-BIO [10], a dataset of biomedical PubMed abstracts annotated with nested named entity mentions and semantic relations. The data covers both English and Russian abstracts. The annotation covers 8 entity types and 14 relation types. At inference time, systems must classify all ordered pairs of entities within each document, predicting one of the 14 relation labels or no relation; the latter is not included in the output.

3.2. Entity Types

The dataset uses the following 8 entity types, summarised together with their semantic scope in the upper half of Table 1. The entity type distribution across training sets is shown in the lower part of Table 1.

Table 1
Entity type definitions and counts in the training sets.

Entity Type	Description	EN Train	RU Train
ANATOMY	Anatomical structures, organs, tissues	889	8,359
CHEM	Chemicals, drugs, compounds	630	4,779
DEVICE	Medical devices and products	29	735
DISO	Diseases, disorders, pathological conditions	1,171	11,177
FINDING	Clinical findings, signs, symptoms	480	4,816
INJURY_POISONING	Injuries and toxic conditions	40	643
LABPROC	Laboratory and diagnostic procedures	220	1,619
PHYS	Physiological processes and functions	402	5,751
Total		**3,861**	**37,879**

As shown in Table 1, DISO and ANATOMY are the dominant entity types in both languages, together accounting for over 50% of all entity mentions. DEVICE and INJURY_POISONING are rare, particularly in English, with 29 and 40 instances respectively, which limits the type-mask constraints that can be learned from the English training set.

3.3. Relation Types and Distribution

The dataset contains 14 semantic relation types, whose counts across all four splits are presented in Table 2.

As Table 2 makes clear, the distribution is highly imbalanced: SUBCLASS_OF is 17x more frequent than APPLIED_TO in the English training set. This makes macro-averaged F1 substantially harder to optimise than micro-averaged F1.

3.4. Dataset Split Statistics

The overall document and relation counts per split are summarised in Table 3.

As can be seen from Table 3, the English training set is notably small at 55 documents, roughly 12x fewer than Russian. Combined with the nested entity annotation, this makes English the hardest track.

3.5. Candidate Pair Explosion at Test Time

At inference time, systems must classify *all* ordered entity pairs in each document, not just those with known relations. This leads to a dramatically larger number of candidate instances at test time compared to training, as shown in Table 4.

Table 2
Relation counts per split.

Relation	EN Train	EN Dev	RU Train	RU Dev
SUBCLASS_OF	743	757	7,517	669
HAS_CAUSE	579	487	2,710	407
AFFECTS	362	393	2,493	288
ASSOCIATED_WITH	325	233	3,295	242
PART_OF	204	248	1,483	208
PHYSIOLOGY_OF	150	178	1,127	131
FINDING_OF	121	83	1,356	85
TO_DETECT_OR_STUDY	164	116	771	84
ORIGINS_FROM	120	79	566	67
ALTERNATIVE_NAME	95	98	649	97
ABBREVIATION	97	74	862	67
TREATED_USING	101	88	282	46
USED_IN	89	79	378	85
APPLIED_TO	43	54	284	45
Total	**3,193**	**2,967**	**23,773**	**2,521**

Table 3
Document and relation counts per split.

Language	Split	Documents	Relations	Avg. rel/doc
English	Train	55	3,193	58.1
English	Dev	50	2,967	59.3
Russian	Train	661	23,773	36.0
Russian	Dev	50	2,521	50.4
Bilingual	Train+Dev	816	32,454	39.8

Table 4
Training vs. test instance counts (including negatives).

Split	Instances	Positive ratio
EN train	12,739	∼25%
EN dev	11,838	∼25%
EN test	818,084	unknown
RU train	95,065	∼25%
RU dev	10,069	∼25%
RU test	664,448	unknown
BIL traindev	129,711	∼25%

As Table 4 shows, the test sets contain roughly 65–70x more candidate instances than the development sets. This comparison is only approximate, however: the training and development instance counts reflect the 3:1 negative-sampling policy applied at training time, which yields the ∼25% positive ratio shown, whereas the test counts enumerate every ordered entity pair with no sampling at all. The figure therefore conflates a change in sampling policy with the genuine growth in the number of document-level candidate pairs. In either reading, all possible entity pairs in every test document are enumerated as candidates without negative sampling, so distance pruning and type masks are critical for managing this scale at inference time.

4. Method

Our system follows a pipeline architecture: (1) candidate pair enumeration from the entity annotations, (2) sentence-level relation classification using a fine-tuned pre-trained language model, and (3) post-processing with type masks, threshold calibration, and distance pruning.

4.1. Model Architecture

Our architecture deliberately follows the official BioNNE-R baseline released by the organisers [1], which couples OpenNRE's [2] `BERTEntityEncoder` with a `SoftmaxNN` classification head and an AdamW optimiser, inserting the markers `[E1]/[/E1]/[E2]/[/E2]` around the head and tail spans. We inherit this encoder–classifier design essentially unchanged and concentrate our contribution on four modifications: (i) replacing the baseline mBERT backbone with a stronger domain- or language-specialised encoder, namely BioLinkBERT-large for English and XLM-RoBERTa-large for Russian and bilingual; (ii) training the final models on the combined train+dev data; (iii) a two-seed logit-averaging ensemble for the English track; and (iv) the three post-processing steps described in Section 4.5. We chose this entity-marker classification design over span-pair or graph-based relation extraction architectures for two reasons: it reproduces a well-understood, strong baseline with minimal implementation risk under the shared-task timeline, and it is comparatively data-efficient, an important consideration given that the English track provides only 55 annotated training documents (Section 3), a regime in which the larger parameter count and structural priors of graph models are difficult to fit reliably.

Concretely, for each candidate pair the encoder constructs the input sequence by inserting entity marker tokens around the head and tail entity spans within the document text:

```
... [E1] head entity [/E1] ... [E2] tail entity [/E2] ...
```

Following the standard entity-marker formulation of the `BERTEntityEncoder`, the transformer hidden states at the two entity-marker start tokens (`[E1]` and `[E2]`) are concatenated and passed through a linear projection to form the candidate-pair representation. This representation is fed to a `SoftmaxNN` classification head with softmax activation over 15 classes (14 relation types + `no_relation`), trained with cross-entropy loss and the AdamW optimiser.

For the English track we use `michiyasunaga/BioLinkBERT-large` [8] as the encoder backbone. BioLinkBERT is pre-trained on PubMed abstracts and additionally exploits inter-document citation links to learn entity co-occurrence across documents, which is beneficial for biomedical relation extraction.

For the Russian and bilingual tracks we use `xlm-roberta-large` [7], a 560M-parameter multilingual transformer trained on 2.5 TB of text in 100 languages including Russian.

4.2. Training Configuration

All hyperparameters are identical across tracks and are listed in Table 5.

Table 5
Training hyperparameters.

Parameter	Value
Max sequence length	256 tokens
Batch size	16
Learning rate	2×10^{-5}
Optimiser	AdamW
Warmup steps	300
Max epochs	10
Checkpoint selection	best micro F1 on dev
Negative sampling ratio	3:1 (`no_relation` : positive)

The negative sampling ratio of 3:1 balances the training signal: each positive instance is accompanied by roughly three negative (`no_relation`) instances, which yields the ∼25% positive rate of the training and development splits reported in Table 4. Without such sampling, the model would see overwhelmingly many negatives and under-predict relations. We do not use an early-stopping patience: each model is trained for the full budget of 10 epochs, and we retain the checkpoint with the highest micro F1 on the development set, following OpenNRE's default best-checkpoint selection.

For the final test submissions, we train on the combined train+development set (*traindev*), using all labelled data. This avoids discarding 2,967 English or 2,521 Russian labelled instances from the development set.

For the English track, we additionally train a second model with a different random seed, seed 1 in addition to seed 42, and average the logits of both models at inference time. This *seed ensemble* reduces variance from the stochastic training process.

4.3. Experimental Setup

All models were trained and evaluated on a single NVIDIA GeForce RTX 5070 Ti GPU (16 GB), using PyTorch with the HuggingFace Transformers library and the OpenNRE toolkit [2]. Training used a fixed budget of 10 epochs at batch size 16 and sequence length 256, which fits comfortably within 16 GB for both backbones. Because the number of optimisation steps is fully determined by the traindev instance counts of Table 4, a single English model with BioLinkBERT-large over about 24.6k traindev instances is trained for roughly 15.4k steps, a Russian model with XLM-RoBERTa-large over 105,134 instances for roughly 65.7k steps, and the bilingual model over 129,711 instances for roughly 81.1k steps; the English track trains two such models for the seed ensemble. At inference the classifier performs a single forward pass over the candidate pairs that survive distance and type pre-filtering; for the English test set this amounts to roughly 147,000 of the 818,084 enumerated pairs (Section 5.3), so the inference cost is dominated by this one-time exhaustive sweep rather than by any per-query computation. The code and calibrated per-class thresholds are available from the author on request.

4.4. Data Preprocessing

A key preprocessing decision concerns how to construct input instances from the raw entity annotations. For each document, we enumerate all ordered pairs of entity mentions (head, tail) and construct one input sequence per pair, following the standard pipeline for sentence-level relation extraction [2]. The ordering is important: the pair (A, B) and the pair (B, A) are treated as distinct instances, since the relation may be directional, as SUBCLASS_OF is not symmetric. For entity mentions with discontinuous spans, for example a span recorded as `472-476,492-500` representing a phrase with a non-entity gap, we use the full extent from the first start position to the last end position as the effective span for marker insertion. While this approximation loses the gap information, it preserves the character-level boundaries that the tokeniser requires and is consistent with how the entity markers are defined in OpenNRE. We did not separately quantify how often a gold relation has at least one discontinuous-span endpoint; in NEREL-BIO such mentions constitute a low single-digit percentage of all mentions, so we expect the effect of this approximation on overall macro F1 to be small, and we leave a precise measurement to future work.

Tokenisation is performed using the tokeniser of the respective backbone model: the BioLinkBERT tokeniser for the English track and the XLM-RoBERTa tokeniser for the Russian and bilingual tracks. In both cases, the four entity markers `[E1]`, `[/E1]`, `[E2]`, `[/E2]` are added to the tokeniser vocabulary as special tokens before fine-tuning, ensuring that they are never split into subword units. Sequences that exceed 256 tokens after marker insertion are truncated at the right boundary. This truncation affects only a small fraction of instances, primarily those involving entities near the end of the longest abstracts, and is therefore unlikely to have a significant impact on overall performance.

Relation labels are mapped to integer indices via a fixed label-to-index mapping (`rel2id.json`) shared across all three tracks. The `no_relation` class is assigned index 0 and is included during

training as the target label for negative instances, but it is excluded from evaluation and never written to the submission files. This design follows the standard convention in OpenNRE and ensures that the model learns to distinguish true relations from background entity co-occurrence, rather than treating every candidate pair as a positive example.

4.5. Post-Processing

A notable property of the BioNNE-R inference regime is that the neural classifier alone is insufficient to produce well-calibrated predictions under the extreme class imbalance of the candidate-pair enumeration described in Section 3.5. We therefore apply three post-processing steps in sequence after the model has produced softmax probabilities for every candidate pair.

The first step exploits a structural constraint inherent to the annotation schema: each relation type is semantically defined over a restricted set of entity type combinations. For example, TREATED_USING connects a disease (DISO), an injury/poisoning (INJURY_POISONING), or a physiological process (PHYS) with a chemical (CHEM), but not, say, two anatomical structures. We operationalise this constraint by collecting, for each of the 14 relation types, the complete set of (head_type, tail_type) combinations observed in the training data. Because a single global mask is used for all three tracks, the sets are taken as the union over the English and Russian training annotations; for the bilingual track, a candidate is therefore admissible whenever its type pair is legal in either language. At inference time, any candidate pair whose entity types do not appear in the allowed set for a given relation is forced to `no_relation`, regardless of the model's predicted probability. Table 6 illustrates a subset of these constraints for six selected relations; the full set of constraints is derived analogously for all 14 relation types. This approach is related to type-constrained decoding in knowledge base completion, where ontological type constraints are used to prune impossible (subject, predicate, object) triples. In our setting, the constraints are learned empirically rather than imposed from an external ontology, which makes them directly consistent with the annotation decisions in NEREL-BIO without requiring any additional resources.

Table 6
Example allowed (head_type, tail_type) pairs per relation (examples observed in the combined English and Russian training data). Complete sets are shown for TREATED_USING, PART_OF, and PHYSIOLOGY_OF.

Relation	Allowed (head, tail) pairs
SUBCLASS_OF	(DISO, DISO), (ANATOMY, ANATOMY), (CHEM, CHEM), (PHYS, PHYS)
HAS_CAUSE	(DISO, CHEM), (DISO, DISO), (FINDING, CHEM), (INJURY_POISONING, CHEM)
ABBREVIATION	(DISO, DISO), (CHEM, CHEM), (LABPROC, LABPROC)
TREATED_USING	(DISO, CHEM), (INJURY_POISONING, CHEM), (PHYS, CHEM)
PART_OF	(ANATOMY, ANATOMY)
PHYSIOLOGY_OF	(PHYS, ANATOMY)

The second step addresses the observation that biomedical relations are predominantly local: they connect entities that co-occur within a short textual window and are typically linked by a single predicate or prepositional phrase. We therefore prune all candidate pairs for which the absolute character distance between the start positions of the head and tail entity spans exceeds 200 characters. This threshold was chosen empirically from the distance distribution on the development set, where the vast majority of annotated relations fall within a 200-character window. Pruning is applied as a pre-filter over the candidate pairs rather than as a rule over already-predicted positives: for the single-model tracks the pruned pairs are removed before the encoder is applied, and for the English seed ensemble the same cutoff is applied to the averaged logits; in both cases pairs beyond 200 characters receive no positive prediction. Long-range pairs that would otherwise survive tend to involve sentence-level co-reference or discourse-level associations rather than direct syntactic relations, and the sentence-level encoder used in our architecture is not well suited to capturing such dependencies. Distance-based candidate filtering of this kind is a standard heuristic in sentence- and window-level biomedical RE, consistent with the general finding that relation extraction performance degrades rapidly as entity distance increases.

The third step addresses a well-known failure mode of softmax classifiers under class imbalance: because the model is trained on a distribution where no_relation constitutes approximately 75% of instances, it learns relatively conservative confidence estimates for rare positive classes. As a result, argmax decoding, which assigns each candidate to its highest-probability class without any minimum confidence requirement, systematically under-predicts relations such as ABBREVIATION and APPLIED_TO, for which the model rarely assigns probability above 0.5 even for true positives. To correct this, we tune a separate decision threshold τ_r for each relation r on the development set. Writing $\hat{r} = \arg\max_c p(c \mid \text{pair})$ for the highest-probability class over all 15 labels, a candidate pair is assigned relation $\hat{r}$ if and only if

$$\hat{r} \neq \texttt{no\_relation} \quad \text{and} \quad p(\hat{r} \mid \text{pair}) \geq \tau_{\hat{r}}. \tag{1}$$

Each threshold τ_r is selected by a grid search over $\{0.05, 0.10, \ldots, 0.95\}$ that maximises the relation's F1 on the development set. The search is a single pass in which each τ_r is optimised independently, with all other relations left at argmax decoding; because a threshold only gates its own relation's argmax-positive predictions in Equation (1), the per-relation optimisations do not interact and the outcome is independent of the order in which relations are visited, and the procedure is not an iterated coordinate descent to convergence. For the English track the thresholds are calibrated on the seed-averaged development logits, the same ensemble used at test time, rather than on either individual seed, and using a train-only model so that the development set remains held out; the resulting values are then reused for the final traindev models. The thresholds obtained for the English ensemble are reported in Table 7 and reveal a clear pattern: rare relations such as ABBREVIATION, APPLIED_TO, and AFFECTS receive low thresholds near 0.05, allowing the model to commit to a prediction at low confidence, while frequent relations such as HAS_CAUSE, ALTERNATIVE_NAME, and PART_OF receive high thresholds of 0.55–0.60, suppressing overconfident false positives. This per-class calibration is closely related to temperature scaling [14] and Platt scaling, but operates directly on the macro F1 objective rather than on log-likelihood, making it better suited to the imbalanced evaluation metric of this task.

Table 7
Calibrated thresholds for the EN ensemble (selected relations).

Relation	Threshold τ_r
ABBREVIATION	0.05
APPLIED_TO	0.05
AFFECTS	0.05
ORIGINS_FROM	0.05
USED_IN	0.05
HAS_CAUSE	0.55
ALTERNATIVE_NAME	0.60
PART_OF	0.60

5. Results and Discussion

5.1. Ablation Study

To understand where the improvement over the baseline originates, we compare the official mBERT baseline with our stronger backbone on the development set, training only on the train split so that the development set stays held out. The two configurations are shown in Table 8. We were not able to log every intermediate configuration under identical conditions, whether each post-processing step in isolation or training on traindev; rather than report imprecise per-component development deltas, we therefore quantify only these two anchors here and defer the combined effect of the remaining choices to the held-out test results in Table 9.

Table 8
Development-set macro F1 for the mBERT baseline and our stronger backbone, before any post-processing. Both configurations are trained on the train split only, keeping the development set held out. The baseline figures are exact; the backbone-swap figures are approximate single-run estimates. Per-component post-processing deltas and the effect of training on traindev are not tabulated here (see text) and are reflected in the held-out test results of Table 9.

System (dev macro F1)	EN Dev	RU Dev
Official baseline (mBERT + OpenNRE)	0.3435	0.4844
+ stronger backbone, before post-processing	$\approx$0.60	$\approx$0.68

The dominant contribution by a wide margin comes from replacing the mBERT backbone with a larger, more specialised encoder: BioLinkBERT-large for English and XLM-RoBERTa-large for Russian and bilingual. This single change raises development macro F1 from 0.3435 to approximately 0.60 on English and from 0.4844 to approximately 0.68 on Russian, gains of roughly 0.26 and 0.20 respectively. The magnitude of this improvement reflects both the larger model capacity of these encoders, which have roughly 2–3x more parameters than mBERT, and the quality of their pre-training corpora: BioLinkBERT is specifically trained on PubMed with citation-link augmentation, while XLM-RoBERTa is trained on a much larger and more diverse multilingual corpus than mBERT.

On top of the stronger backbone, the three post-processing steps further improve performance; because we did not log each step in isolation under identical conditions, we describe their effect qualitatively rather than as precise development deltas. Per-class threshold calibration had the clearest effect, consistent with the hypothesis that the uncalibrated softmax classifier systematically misallocates confidence across relation types under the 3:1 negative sampling regime; this effect is most pronounced on English, whose small development set leaves the default argmax thresholds particularly poorly tuned for rare relations. Entity-type masks provided a smaller but consistent gain by removing type-implausible false positives; this gain was larger on English, where the model produces more such errors, likely because it has seen fewer examples per relation type during training. Distance pruning contributed a further small improvement by discarding long-range spurious pairs.

Two further choices define the submitted system but are deliberately excluded from the comparison above, because neither can be measured as a clean held-out ablation step. Training on the full traindev set folds the development data into training, so any development score for it would be an in-sample estimate rather than a generalisation gain; and for English we additionally average two seeds at inference time. The effect of both choices is therefore visible directly in the held-out test results of Table 9, where the seed ensemble in particular accounts for roughly +0.011 macro F1 on English (Section 5.3).

5.2. Test Results

The final test set scores for all three tracks are presented in Table 9.

Table 9
Test set results (macro F1) compared to the official baseline.

Track	Baseline	Our system	Δ
English	0.2825	**0.4570**	+0.1745
Russian	0.3070	**0.4829**	+0.1759
Bilingual	0.3027	**0.4849**	+0.1822

As shown in Table 9, our system outperforms the official OpenNRE mBERT baseline by more than 0.17 macro F1 on all three tracks. The absolute gains are nearly identical across tracks, at +0.1745, +0.1759, and +0.1822, suggesting that our post-processing pipeline generalises well regardless of language and model backbone. The bilingual model achieves the highest score despite being a single model required to handle two languages simultaneously. This is most plausibly explained by training data

volume: the bilingual traindev set contains 129,711 instances, compared to 105,134 for the monolingual Russian traindev, which combines 95,065 training and 10,069 development instances, a 1.2x advantage in candidate pairs, though proportionally more positive relation instances due to the added English annotations. The fact that the bilingual model also slightly outperforms the monolingual Russian model, 0.4849 against 0.4829, suggests that exposure to English data during training may act as a form of regularisation or data augmentation for the Russian track, a phenomenon consistent with findings on cross-lingual transfer in multilingual transformers [7].

We caution, however, that the reported Russian and bilingual test scores are single-run point estimates obtained with seed 42; only the English result is a two-seed logit average. After the second seed collapsed to a degenerate solution on the bilingual track (Section 5.3), we did not have the compute budget to obtain a multi-seed mean $\pm$ standard deviation for the Russian and bilingual tracks. Given the seed sensitivity we document, the 0.002 macro-F1 gap between the bilingual and monolingual Russian systems is within the plausible range of run-to-run variation, so the claim that the bilingual model outperforms the Russian one should be read as suggestive rather than conclusive.

Beyond the organiser baseline, it is informative to situate our runs against the other participating systems. According to the shared-task overview [1], our submissions placed within the top 6 on the English track, within the top 3 on the Russian track, and second on the bilingual track, confirming that the gains reported in Table 9 are competitive with the strongest systems in the shared task rather than an artefact of a weak baseline.

5.3. Discussion

An interesting observation from Table 9 is that the English system, at 0.4570, underperforms the Russian one, at 0.4829, despite relying on the domain-specifically pre-trained BioLinkBERT-large backbone, which is generally regarded as a stronger starting point for English biomedical NLP than XLM-RoBERTa-large. We argue that this apparent paradox is explained by the severe training data imbalance between the two languages: the English portion of NEREL-BIO contains only 55 annotated documents, whereas the Russian portion contains 661, a 12x difference.

With so few English training examples, the type-mask constraints derived from training data are incomplete: several legal type combinations never appear in the 55-document English training set, and rare entity types such as DEVICE and INJURY_POISONING appear too infrequently to yield reliable allowed (head_type, tail_type) statistics. Similarly, per-class threshold calibration becomes noisier when the development set is small: even a handful of prediction changes can shift the estimated optimal threshold by a full grid step, 0.05 in our setup, translating into a non-trivial change in macro F1 for rare relations. This data-scarcity effect compounds a well-documented tendency for the fine-tuning of large pre-trained encoders to be unstable and high-variance across random seeds [13]. Domain-specific pre-training, as in BioBERT [5] and PubMedBERT [6], improves absolute performance on standard-sized biomedical benchmarks, but such advantages are known to shrink and become less reliable when only a few dozen labelled documents are available. Our own results are consistent with this pattern: the backbone advantage of BioLinkBERT-large is real, but it cannot compensate for a 12x gap in supervision. The most promising remedy would be to obtain additional English annotations or to use cross-lingual projection from the richer Russian training data.

A second notable finding concerns the effect of the seed ensemble on the English track. Averaging the logits of two models trained with different random seeds, seed 42 and seed 1, improved test macro F1 by approximately 0.011. This gain is modest in absolute terms but consistent with the variance reduction interpretation of ensembling: when the training set is small, individual fine-tuning runs explore different regions of the loss landscape and converge to solutions with partially complementary strengths and weaknesses [13]. The magnitude of the improvement we observe is in line with results reported for 2-model seed ensembles in other low-resource classification tasks. We attempted to apply the same strategy to the bilingual track, but encountered a qualitatively different outcome: the second seed produced a degenerate model that achieved micro F1 of approximately 0.10 on the development set, converging to a solution that predicted nearly all instances as the dominant class. Fine-tuning runs

that collapse or sharply underperform under particular random seeds have been documented for large pre-trained encoders [13]; in our case the collapse coincided with the aggressive negative sampling and heavy class imbalance of the bilingual setting. The bilingual data mixture, which combines Russian and English in roughly equal proportion, likely introduces additional gradient instability during the early training epochs, before the model has established language-agnostic representations. Addressing this through more carefully tuned warm-up schedules or language-balanced sampling strategies could reduce the probability of such collapses in future work.

Finally, a structural challenge that shaped the design of our entire post-processing pipeline is what we term the *candidate pair explosion*: the test set requires classification of all ordered entity pairs within each document, resulting in 818,084 candidates for English and 664,448 for Russian, roughly a 65–70x increase relative to the negatively-sampled development set, as previously noted in Table 4. This increase arises from the evaluation protocol rather than from any property of the test documents themselves. The consequence is a severely imbalanced inference regime: our system outputs 19,988 positive predictions from 818,084 English candidates, a positive rate of approximately 2.4% (see Table 10 for the per-relation breakdown). Maintaining adequate precision under such extreme imbalance is a fundamental challenge for exhaustive-pair relation extraction: even a classifier with 99% specificity would generate over 8,000 false positives on the English test alone. The two candidate filters act as strong pre-filters here. On the English test set, the distance cutoff alone retains 24.0% of the 818,084 candidate pairs, eliminating 76.0%, and the entity-type mask alone retains 73.1%, eliminating 26.9%; applied together they retain only 18.0% of all candidate pairs, about 147,000, which the classifier and per-class thresholds then reduce to the 19,988 emitted positives. Nevertheless, threshold calibration alone is insufficient to fully resolve the precision–recall tension at this imbalance level, and a dedicated two-stage pipeline, in which a lightweight binary filter first reduces the candidate set and the full neural classifier is applied only to surviving pairs, represents a natural next step.

Table 10
Final prediction counts by relation (test set).

Relation	EN Test	RU Test
HAS_CAUSE	4,790	2,072
SUBCLASS_OF	3,761	3,496
AFFECTS	2,695	1,526
ASSOCIATED_WITH	2,403	1,967
PART_OF	1,007	871
PHYSIOLOGY_OF	934	696
TO_DETECT_OR_STUDY	861	751
FINDING_OF	725	641
ALTERNATIVE_NAME	640	477
ABBREVIATION	631	381
USED_IN	614	313
ORIGINS_FROM	476	401
TREATED_USING	280	162
APPLIED_TO	171	109
Total	**19,988**	**13,863**

6. Conclusion

We presented a system for biomedical nested relation extraction in the BioNNE-R shared task at BioASQ 2026. The system extends the OpenNRE framework with three task-specific post-processing components: entity-type masks, distance-based candidate pruning, and per-class threshold calibration. These components are applied on top of strong pre-trained language model backbones, namely BioLinkBERT-large for the English track and XLM-RoBERTa-large for the Russian and bilingual tracks.

The final system achieves macro F1 scores of 0.4570 for English, 0.4829 for Russian, and 0.4849 for the bilingual track, outperforming the official mBERT-based OpenNRE baseline by more than 0.17 macro F1 in all three settings. The ablation results suggest that the largest improvement comes from replacing the baseline encoder with a stronger pre-trained backbone, while per-class threshold calibration provides the largest gain among the post-processing components. Type masks and distance pruning provide smaller but consistent improvements by reducing implausible positive predictions under the highly imbalanced candidate-pair regime.

These results indicate that a sentence-level relation extraction pipeline, when combined with carefully designed task-specific post-processing, can serve as an effective and easily reproducible baseline for nested biomedical relation extraction. At the same time, the results also show the limitations of this approach: by treating each ordered pair independently, it forgoes any explicit modelling of the nested entity structure that a span-pair or graph-based architecture would provide. The English track remains challenging because of the small amount of training data, which restricts the completeness of type constraints and makes threshold calibration less stable for rare relation types. We also observed training instability in one bilingual run, suggesting that large multilingual transformers can be sensitive to random seed choice under aggressive negative sampling and strong class imbalance.

Future work should address these limitations in three directions. First, cross-lingual transfer from the larger Russian portion of the dataset could be used to improve English performance, for example through language-balanced training or translation-based augmentation. Second, graph-based or span-pair architectures could explicitly model the nested structure of entity mentions instead of treating each ordered pair independently. Third, more robust calibration and candidate filtering methods could be explored, including binary candidate pre-filtering, relation-specific negative sampling, and data augmentation for low-resource relation types such as APPLIED_TO and ABBREVIATION.

Acknowledgments

The author thanks the organisers of the BioNNE-R shared task at BioASQ 2026 for providing the data and evaluation infrastructure, and the NEREL-BIO annotation team for making the dataset publicly available.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools solely as writing assistants, under full human supervision, for language editing, formatting assistance, and proofreading. Specifically, a large language model assistant was used to improve grammar, spelling, and phrasing. After using these tools, the author reviewed and edited all content and takes full responsibility for the publication's content.

References

[1] Rozhkov, I., Loukachevitch, N., Tutubalina, E.: Overview of the BioASQ BioNNE-R Task on Nested Biomedical Relation Extraction at CLEF 2026. In: CLEF 2026 Working Notes (2026).

[2] Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., Sun, M.: OpenNRE: An Open and Extensible Toolkit for Neural Relation Extraction. In: Proceedings of EMNLP-IJCNLP 2019 (System Demonstrations), pp. 169–174 (2019).

[3] Krallinger, M., Rabal, O., Akhondi, S. A., et al.: Overview of the BioCreative VI chemical-protein interaction Track. In: Proceedings of the Sixth BioCreative Challenge Evaluation Workshop, Vol. 1, pp. 141–146 (2017).

[4] Luo, L., Lai, P. T., Wei, C. H., Arighi, C. N., Lu, Z.: BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 23(5), bbac282 (2022).

[5] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., Kang, J.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234–1240 (2020).
[6] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23 (2021).
[7] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-lingual Representation Learning at Scale. In: Proceedings of ACL 2020, pp. 8440–8451 (2020).
[8] Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: Pretraining Language Models with Document Links. In: Proceedings of ACL 2022, pp. 8003–8016 (2022).
[9] Sakhovskiy, A., Loukachevitch, N., Tutubalina, E.: Overview of the BioASQ BioNNE-L Task on Biomedical Nested Entity Linking in CLEF 2025. CLEF 2025 Working Notes, CEUR Workshop Proceedings, vol. 4038, pp. 41–50 (2025).
[10] Loukachevitch, N., Sakhovskiy, A., Tutubalina, E.: Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian. In: Proceedings of LREC-COLING 2024, pp. 2383–2389 (2024).
[11] Nentidis, A., Katsimpras, G., Krithara, A., et al.: Overview of BioASQ 2024: The twelfth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction (Proceedings of CLEF 2024). Springer LNCS, vol. 14958, pp. 3–27 (2024).
[12] Davydova, V., Loukachevitch, N., Tutubalina, E.: Overview of BioNNE Task on Biomedical Nested Named Entity Recognition at BioASQ 2024. CLEF 2024 Working Notes, CEUR Workshop Proceedings, vol. 3740, pp. 28–34 (2024). URL: https://ceur-ws.org/Vol-3740/paper-03.pdf
[13] Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H., Smith, N. A.: Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. arXiv preprint arXiv:2002.06305 (2020).
[14] Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q.: On Calibration of Modern Neural Networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), pp. 1321–1330 (2017).
[15] Nentidis, A., Katsimpras, G., Krithara, A., Krallinger, M., Rodríguez-Ortega, M., Rodriguez-López, E., Loukachevitch, N., Rozhkov, I., Tutubalina, E., Dimitriadis, D., Patsiou, V., Tsoumakas, G., Giannakoulas, G., Bekiaridou, A., Samaras, A., Di Nunzio, G. M., Ferro, N., Marchesin, S., Martinelli, M., Silvello, G., Paliouras, G.: Overview of BioASQ 2026: The Fourteenth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026). Lecture Notes in Computer Science, Springer (2026).